Preserving a combat commander's moral agency: The Vincennes Incident as a Chinese Room

Patrick Chisan Hew
Defence Science and Technology Group, Department of Defence (Australia)

Abstract: We argue that a command and control system can undermine a commander's moral agency if it causes him/her to process information in a purely syntactic manner, or if it precludes him/her from ascertaining the truth of that information. Our case is based on the resemblance between a commander's circumstances and the protagonist in Searle's Chinese Room, together with a careful reading of Aristotle's notions of 'compulsory' and 'ignorance'. We further substantiate our case by considering the Vincennes Incident, when the crew of a warship mistakenly shot down a civilian airliner. To support a combat commander's moral agency, designers should strive for systems that help commanders and command teams to think and manipulate information at the level of meaning. 'Down conversions' of information from meaning to symbols must be adequately recovered by 'up conversions', and commanders must be able to check that their sensors are working and are being used correctly. Meanwhile ethicists should establish a mechanism that tracks the potential moral implications of choices in a system's design and intended operation. Finally we highlight a gap in normative ethics, in that we have ways to deny moral agency, but not to affirm it.

Keywords: moral agency; Chinese Room; command and control; Vincennes Incident.

1 Introduction

Under the principle of legitimacy, the armed forces of the United States will seek to 'maintain legal and moral authority in the conduct of operations' (Department of Defense (U.S.), 2011). Similar statements can be made about other military forces worldwide. We assert that moral authority in the large may require moral authority of individuals, especially in commanders. We pose the question: How could the design and construction of a command and control system undermine a commander's moral agency?

To expose such deficiencies, we pursue the resemblance between a commander's circumstances and the protagonist in Searle's Chinese Room. The resemblance has yet to be examined carefully in the literature, and leads to the following insights:

1. *Moral agency can be undermined when the system causes a commander and/or command team to process information in a purely syntactic manner, or if it precludes them from ascertaining the truth of that information*. Designers should therefore strive for systems that help commanders and command teams to think and manipulate information at the level of meaning, even when circumstances are driving otherwise. While the desirability of such thinking can be argued on other merits, we make a novel case from the perspective of normative ethics.
2. *Ethicists should establish a mechanism that tracks the potential moral implications of choices in a system's design and intended operation*. We will identify two choices that can undermine a commander's moral agency. Ethicists need to be able to track such choices.

3. *If a commander's moral agency is undermined then the responsibilities that were held by the commander would need to be reapportioned, but how this ought to be done is unclear.* There appears to be a gap in normative ethics, in that we have ways to deny moral agency, but not to affirm it.

The paper proceeds as follows: We start with our analysis of commanders, their teams and their systems, and the resemblance with the Chinese Room. We deduce some necessary conditions for moral agency. We test our proposal by examining the shooting-down of Iran Air Flight 655 by the warship *USS Vincennes* (the Vincennes Incident). We round out with conclusions.

2 Commanders in the Chinese Room

2.1 The Chinese Room Argument

Searle (1980) posed his Chinese Room Argument as a rebuttal to the hypothesis of Strong Artificial Intelligence (AI). According to Searle, Strong AI proposed that an appropriately programmed computer with the right inputs and outputs would have a mind in exactly the same sense that human beings have minds. Weak AI, by contrast, merely claims that an appropriately programmed computer can simulate mind. The hypothesis of Strong AI (as defined by Searle) concurred with then-contemporary claims by early AI researchers (Cole, 2015).

The Chinese Room Argument frames a thought experiment: John is seated in a room, fully-enclosed on all sides. He is equipped with baskets of cards, each card holding a Chinese symbol. He also sees a slot from the outside world ('input slot'),

through which cards can enter, and he can place cards into a corresponding slot to the outside world ('output slot').

John does not understand Chinese. He is, however, equipped with a rulebook in a language that he does understand (English). The rulebook tells him the cards that he should put into the output, given the cards that he has received as input. Thus if he sees 'squiggle squiggle' in, he looks up the rulebook, follows its instructions, and generates 'squoggle squoggle' out.

John is in the same situation as a computer that is executing a program. As per the hypothesis of Strong AI, we suppose that the rulebook constitutes an 'appropriate program' in the sense that John can generate the right output to any given input. But John does not understand Chinese. According to Searle, this contradicts the proposition that John-as-computer has mind.

2.2 Application to commanders in combat operations

For this article, we accept the position that John-as-computer lacks mind. We now explore the resemblance between John in the Chinese Room and commanders in combat operations.

We replace John with a *commander*, an human who is properly designated to exercise authority and direction over assigned and attached forces in the accomplishment of a mission (He/she exercises *command and control*, (Department of Defense (U.S.), 2013)). The commander is situated in a room, their *command center*, along with more humans who comprise his/her *command team*. We will continue to use *command and control system* to refer to the command center and all

enclosed components. Real-world examples include a headquarters on land, or the operations room or Combat Information Center aboard a warship (Figure 1).

We replace the input slot and Chinese symbols with a *symbolic representation of the battlespace*. Military standard MIL-STD-2525 (Department of Defense (U.S.), 2014) exemplifies the symbols that could be used. While the symbols could be drawn onto a sheet of plastic overlaid onto a paper map, we focus on electronic presentations from a command and control system. The output slot corresponds to the *control panel* to that system. The setup is a good model of command and control for combat operations, and for air combat and missile defense in particular.

We now contemplate a commander who processes the symbols in a purely syntactic manner, just as John processed Chinese symbols. For example, consider the statement 'If ⌂ moves *inside* the circle, then set flag to true'. The symbols ⌂, circle, flag, true and the relationship of being *inside* are purely syntactic, and the statement constitutes a rule for processing those symbols in a purely syntactic manner.

Having accepted that John-as-computer lacks mind, then we would say that a commander who processes the symbols in a purely syntactic manner will also lack mind. In combat operations, it is unlikely that we would see a commander referring to a literal rulebook. Nonetheless, the conclusion holds if the commander memorizes the rules and implements them by rote. It holds even more strongly if the commander implements the rules out of habit – 'mindlessly'.

2.3 Symbolic representations vs extended human sensing of the battlespace

In analogizing a commander center to a Chinese Room, we replaced the input slot and Chinese symbols with a symbolic representation of the battlespace. We provide details on how the representation is generated in real life. For our purposes, we may regard *sensors* as electro-optical / mechanical devices that measure an aspect of the ambient environment. We then partition into two sub-classes: *extended human sensors* versus *object-inferencing sensors*.

Extended human sensors extend the range of a human's organic sensors. A given sensors' measurements are conveyed to the human as if they were at the sensor's location. Video cameras are a prototypical example – they periodically measure the ambient light, and present it to a human as if that light had come into his/her eyes.

Object-inferencing sensors apply algorithms to the measurements to infer the existence and properties of objects. Radar is a prototypical example – it illuminates the environment with radiofrequency energy, and infers that an object exists ('detects' an object) if the backscatter is sufficiently concentrated in space, time and frequency. An extended human sensor can be turned into an object-inferencing sensor by equipping it with an appropriate algorithm. An example is the face-detection software in a digital camera. Ideally the inferences will be a perfect match with reality, but erroneous inferences can occur.

We restrict our attention to object-inferencing sensors. They are especially prevalent in aerospace operations, since radar can sense to greater distances than

electro-optical cameras. Object-inferencing sensors inevitably lead to a symbolic representation of the battlespace, as the symbols depict the inferences.

3 Necessary condition for commanders to have moral agency

We have asserted that a commander's moral agency is undermined if the command and control system causes them to process information in a purely syntactic manner. We now fill out this claim, to establish some necessary conditions for preserving a commander's moral agency.

3.1 Definitions

(The first four paragraphs of this section are taken from the author's previous work at(Hew, 2014). We are grateful for receiving permission.) We accept an *agent* as something that can act in the world. We declare that an agent is *morally praiseworthy*, and can be held *morally responsible* for an action, if it is worthy of praise for having performed the action. We call such an agent a *moral agent*. (For brevity, 'praise' will cover both praise or blame throughout this article.)

Our choices in terminology sit within what Himma (2009) dubbed the *standard view* of moral agency, citing (Eshleman, 1999; Haksar, 1998; Williams, 2014) as references. Neither he nor we claim that the standard view is correct. We merely seek terminology that has sufficient precision to argue a position, and a starting point of relevance to moral philosophy in the large. Following (Eshleman, 1999), we note a general alignment with Aristotle's *Nicomachean Ethics* Book III.1-5 (Aristotle (translated by W. D. Ross)):

> *Since virtue is concerned with passions and actions, and on voluntary passions and actions praise and blame are bestowed, on those that are involuntary pardon, and sometimes also pity*, to distinguish the voluntary and the involuntary is presumably necessary for those who are studying the nature of virtue, … . Those things, then, are thought involuntary, which take place *under compulsion* or *owing to ignorance*; … . [Book III.1 *emphasis added*]

Aristotle (apparently) makes two proposals. First, that 'virtue' (moral responsibility) is about the bestowing of praise. We accept this proposal as being uncontentious.

His second proposal is that to qualify for such praise, an agent must be capable of 'voluntary' action. We accept this proposal, with the caveat that the detailed nature of 'voluntary' action is still a matter for debate. We will offer a reading of 'compulsory' action and the state of being 'ignorant' that resembles our position. In offering an interpretation of our position with respect to Aristotle's writings, we merely seek to show where our contribution departs from previous studies. We do not claim Aristotle as authority, or that he would have also arrived at our position (we do not 'put words into his mouth').

3.2 Commanders must process information at the level of meaning

We propose our first condition: if a commander is to be a moral agent, then he/she must process information at the level of meaning. As a corollary in converse, if a

system causes a commander to process information in a purely syntactic manner, then his/her moral agency will be undermined.

Our position follows immediately from accepting that John-as-computer lacks mind. Namely, we take moral agency as being a stronger condition than having mind, so a lack of mind implies an absence of moral agency. Or alternately, if John-as-computer can be said to have moral agency, then we should also impute moral agency onto mousetraps, toilet tank-fill valves, thermostats, automobile cruise controls; indeed, anything that mindlessly processes from input to output.

Note that we only seek to *deny* moral agency: if a system causes a commander to process information in a purely syntactic manner, then we deny his/her claim to moral agency. The goal for systems designers is to avoid triggering the denial. In other words, we are not trying to define a sufficient condition(s) for moral agency. Said task appears to be intractable at this time. We only set a necessary condition, but failing to meet that condition is grounds for denying moral agency.

We dwell briefly on the following question for completeness: If a commander loses moral agency under the conditions that we have posed, then who is responsible? Responsibility could potentially be apportioned to those humans who placed the commander in that situation, for example from designing the system, authorising it, building it and so on. To the author's knowledge however, there is currently no established procedure for apportioning the responsibility. The nature of this procedure, and how it is deduced from principles as a matter of normative ethics, appear to be open questions (see also (Hew, 2014) for a similar conclusion regarding responsibility for automated systems).

Our proposal can be compared Aristotle's ponderings on the nature of 'compulsory' action:

> …. that is compulsory of which *the moving principle is outside, being a principle in which nothing is contributed by the person who is acting* … , e.g. if he were to be carried somewhere by a wind, or by men who had him in their power. [Book III.1 *emphasis added*]
>
> *What sort of acts, then, should be called compulsory? We answer that without qualification actions are so when the cause is in the external circumstances and the agent contributes nothing*. … . [Book III.4 *emphasis added*]

In the first block of emphasized text, what Aristotle called the 'moving principle' is what we would call the rulebook that was supplied to John. In our reading, John acts in the sense of manipulating cards from input slot to output slot in accordance with that rulebook. However he contributed nothing to that rulebook.

In the second block of emphasized text, we read 'cause' in the sense of 'explanation', or 'the reason for something happening'. The reading concurs with translations from Greek that are nearest to contemporary usages in English (Soccio, 2009). Thus the 'cause' (explanation) for the Chinese Room's outputs is to be found in the rulebook, to which John contributed nothing.

(It is tempting to make further assignments of 'cause' in this vein; for example, to describe John as being the 'material cause', the rulebook as the 'efficient cause' and so on. We refrain from doing so on the understanding that Aristotle's 'four causes' were developed primarily for the study of nature (Falcon, 2015). While they could be

applied to other subjects – they are not exclusively for the study of nature – definitively applying them to our situation needs more care then we are able to devote here.)

3.3 Commanders must ascertain that the symbols represent the truth

We infer a second condition: if a commander is to be a moral agent and he/she is working from a symbolic representation of the battlespace, then he/she must be able to ascertain that the representation is truthful. Likewise, he/she must be able to ascertain that the symbols that he/she makes as outputs are a truthful representation of the actual consequences. In the lexicon of studies concerning the Chinese Room Argument, he/she must *ground* the symbols in reality (adopting terminology from the Symbol Grounding Problem, which was reformulated out of the Chinese Room Argument by Harnad (1990)).

The designers of the command and control system assemble object-inferencing sensors and electronic displays to represent the battlespace symbolically. If they decline to equip the commander so that he/she can ascertain that the system is working correctly, then they preclude his/her moral agency. Indeed the representation of the battlespace depicts the sensors' inferences, but the truth of those inferences depends on the sensors' machinery and usage. If the commander is to have moral agency, then a necessary (possibly insufficient) requirement is that he/she can validate that the sensor is in working order, and is being used in a manner that is believed (to high confidence) to give correct results. So if it is being used under particular

atmospheric conditions for example, then that usage should have previously been cleared for use through modelling, simulation, field trials and the like.

While the requirement can posited as a result of thinking about the Chinese Room Argument, we do not claim it as a direct deduction. The Chinese Room Argument considers John as processing symbols in a purely syntactic manner. In contrast, we are now supposing that John understands the information that he sees (it is presented in English say), but he lacks the means to validate it.

Our proposal can be compared with Aristotle's thoughts concerning 'ignorance'

> Indeed, we *punish a man for his very ignorance, if he is thought responsible for the ignorance*, as when penalties are doubled in the case of drunkenness; for the moving principle is in the man himself, since he had the power of not getting drunk and his getting drunk was the cause of his ignorance. And we punish those who are ignorant of anything … that they are thought to be ignorant of through carelessness; *we assume that it is in their power not to be ignorant, since they have the power of taking care*. [Book III.5 *emphasis added*]

For our purposes, 'ignorance' refers to the commander's perception of reality as imputed from the symbols. The ability to take care operates in two aspects: first, in having the representation available to them, and in the means for validating the truth of that representation. So if the commander assigns meanings to the symbols that diverge from reality, then that is a positive choice to be 'ignorant'. He/she is accountable for making that choice. And if the commander has the means to validate the sensors but chooses not to, then that is a positive choice to be 'ignorant'.

4 Case study – the Vincennes Incident

The Vincennes Incident refers to the shooting down of Iran Air Flight 655 by the United States Navy cruiser USS *Vincennes* on 3 July 1988. Iran Air Flight 655 was a civilian Airbus airliner, and 290 passengers and crew were killed. The shooting-down was quickly acknowledged by the US Navy as something that should not have occurred, and a matter of utmost concern. At the time, the USS *Vincennes* was one of the US Navy's most modern guided-missile cruisers. She was equipped with the state-of-the-art in sensors, weapons and command systems for anti-air warfare, and her crew were highly trained and motivated.

The Vincennes Incident still looms large in in cognitive engineering, having galvanized a generation of research with the aim of preventing a recurrence. Previous analyses have considered the conditions leading into the Incident, the design and construction of systems, the actions taken by the people and the results. To the extent that the analyses took a moral position, things that led to bad outcomes were regarded as morally wrong. We characterize such analyses as following a consequentialist approach.

We add a normative perspective: during the Incident, was the commander processing information at the level of meaning? Our ultimate goal is to motivate and inform the design of future command and control systems. In the analysis of command and control systems to inform design, a purely consequentialist approach can be vulnerable to criticisms of being based on rare and extreme incidents (such as the Vincennes Incident). A purely normative approach can be vulnerable to criticisms

of being divorced from reality. Adding our normative perspective strengthens the consequentialist analyses and is in turn strengthened by them.

The literature on the Vincennes Incident is vast. Our primary source is the official investigation by Fogarty (1988), which is generally accepted as an accurate history. We draw on Klein (1996) for his cogent summary of the events, his listing of 15 key problems that led to the shoot-down, and his commentary on previous theories. Klein in turn refers to Roberts and Dotterway (1995) for details that will interest us, and we supplement with Dotterway (1988). In turn, Roberts and Dotterway cite recollections by Captain Will Rogers (Rogers, Rogers, & Gregston, 1992), the commanding officer of *Vincennes* during the incident. We refer to Polmar (2001) for technical information about *Vincennes* and her systems.

We initially hypothesized that the design of *Vincennes's* Combat Information Center caused the commander and command team to process information in a purely syntactic manner. Actually diagnosing such an occurrence is extremely difficult as a matter of principle, in that it amounts to solving the problem of other minds (!). We will return to this point later in discussing 'story generation strategies' used by experienced decision makers.

We instead found ourselves with a more interesting hypothesis: that the transactions from human to human, human to machine and/or machine to human caused some of that information to be processed syntactically. Information was 'down converted' from meaning to symbols so that it could be transacted, and inadequately 'up converted' from symbol to meaning. In this light, we focus on three problems

identified by Klein (1996), denoted by him as Difficulties #3, #11 and #12. Klein judged these problems as having a high impact on the incident.

Our first example (Difficulty #3) centers on the UPX-29 Interrogation Friend or Foe (IFF) system. IFF illuminates a volume of airspace with a coded signal. Aircraft within that volume respond with a counter-signal (or 'squawk') providing information including course, speed, altitude and identity. By contrast, radar illuminates a volume of airspace with radiation, but aircraft in that volume do not respond (indeed they may lack the means to detect the illumination). The radar infers the aircraft's properties by processing radiation that has been backscattered.

Iran Air Flight 655 departed Bandar Abbas airport at 1017 local time. It was detected and tracked by *Vincennes*'s radar, and the crew interrogated it with IFF. The Airbus responded with a Mode III squawk, a code generally associated with a civilian aircraft for air traffic control purposes. Its flight path put it on a course to close with the *Vincennes*.

At 1020, the UPX-29 reported a Mode II squawk. At the time of *Vincennes*'s deployment, IFF Mode II was regularly used by Iranian military aircraft, so the crew classified the inbound contact accordingly. The UPX-29 was, however, not actually interrogating the Airbus. The IFF display had been 'hooked' (configured) to display the IFF responses next to the contact's symbol. This gave the impression that the IFF responses should be associated with the contact. While never officially confirmed, plausible speculation (Klein, 1996) holds that the IFF was still interrogating the vicinity of Bandar Abbas airport. Bandar Abbas was a joint military-civilian facility, with any number of Iranian military aircraft.

In effect (if we accept the speculation as per Klein (1996)), the IFF was interrogating the wrong aircraft. The radar was generating symbols for an inbound contact. The IFF was generating symbols for an Iranian military aircraft in the vicinity of Bandar Abbas airport. The conjoined symbols depicted an inbound Iranian military aircraft, but this was a false depiction of the battlespace.

The conjoining of the IFF responses with the radar returns was a syntactic process – a concatenation of symbols that were close to each other *in the representation*. The actual aircraft being represented were at different locations in reality.

Our second example (Difficulty #11 and #12) concerns the confusion of the Airbus with another aircraft. When Iran Air Flight 655 was first detected by the *Vincennes*, it was assigned a track number TN 4474. *Vincennes* was in company with the frigate *USS Sides*, which was also tracking the contact under TN 4131. The tracks were merged, adopting the label TN 4131.

At 1022, the Airbus had closed to within 20 nautical miles of the *Vincennes* and was being tracked as an 'unknown-assumed hostile'. The Commanding Officer asked 'What is 4474 doing?' (Rogers et al., 1992). Unfortunately, track number TN 4474 had been reassigned to another aircraft, at that time descending and accelerating over the Gulf of Oman (Roberts & Dotterway, 1995), (Dotterway, 1988). (Computer scientists will recognize that 'TN 4474' had become a dangling pointer, with dangers that are well known – they are difficult to detect as they rarely trigger an immediate crash, but subsequent computations have unpredictable results.)

Critically, in responding to the request 'What is 4474 doing?', it seems that at least one crew member acted on the 'TN 4474' syntactically, as in literally keying '4-4-7-

4' to retrieve information (Rogers et al., 1992). Looking up 'TN 4474' returned an aircraft that was descending and accelerating. The arguable consequence was that the crew perceived a contact that was unidentified, closing and descending.

We emphasize the way that 'TN 4474' was processed syntactically. At the level of meaning, the Commanding Officer likely meant 'What is the inbound contact doing?'

5 Comparison with earlier work

By reasoning from analogy with the Chinese Room Argument and from first principles, we concluded that if a system causes a commander to process information in a purely syntactic manner, then his/her moral agency will be undermined; likewise if a commander is precluded from validating the system's sensors then his/her moral agency will be undermined. From reanalysing the Vincennes Incident, we further highlight that information may be 'down converted' from meaning to symbols to be transacted between people and/or machines, but inadequately 'up converted' from symbol to meaning.

Our conclusions concur with earlier positions that were arrived at in cognitive ergonomics, where we have added the motivation of preserving a commander's moral agency. In the influential model of situation awareness due to Endsley (1995), we are pointing to qualitatively high levels of situation awareness (so-called Level 3 situation awareness). We also have a strong connection to 'story generation strategies'; as related by Klein (1996): A contact might exhibit the features of a hostile aircraft. But if the decision maker cannot generate a plausible story for the presence of such an aircraft then they will reject the hypothesis. In our terms, if a decision maker is

presented with symbols but cannot ground them with a plausible meaning, they will reject the symbols.

Again within cognitive ergonomics, the consideration of object-inferencing sensors invites further development. Mainstream theory (Endsley, 1995) accepts that inferences can be made by sensors and the system, and then focusses on how they are best conveyed to a human operator. Object-inferencing sensors can be modeled as automation applied to information processing, using for example the 4-stage model by Parasuraman, Sheridan, and Wickens (2000). The implications regarding the responsibilities for that automation have yet to be fully developed, from both a work perspective (ergonomics) and an ethical perspective (moral agency).

Our study of commanders resembles the one by Royakkers and van Est (2010) of 'cubicle warriors', namely people who remotely-controlled military robots from behind a visual interface. They argued that 'cubicle warriors' cannot reasonably (their term) be held responsible for their decisions and actions; in our terminology, that 'cubicle warriors' could not be regarded as moral agents. Key reasons proposed were that the 'cubicle warrior' is detached from reality, and presented with an overly 'clean' picture of the situation on their screen. We concur with their position: if a 'cubicle warrior' operates solely from a symbolic representation of the battlespace, and the meaning of those symbols has been suppressed, then their moral agency is undermined.

Our work may otherwise be regarded as a critical examination of the Moral Turing Test. To summarize, Allen, Varner, and Zinser (2000) proposed that a robot may be regarded as a moral agent if its behaviours are functionally indistinguishable from a

moral person. The Moral Turing Test was subsequently disavowed as a criterion for genuine moral agency, in recognition of controversies surrounding the Turing Test (Allen, Smit, & Wallach, 2005). Indeed we posited our commander (or John-as-computer) generating behaviours that are indistinguishable from a moral person, but not being a moral agent. In this respect, we echo the points made by Bringsjord, Bello, and Ferrucci (2001) – proponents of a Moral Turing Test must articulate why the robot holds moral responsibility, and not its programmer.

6 Conclusion

To support a their moral agency, commanders and command teams need systems that help them to think, and manipulate information, at the level of meaning. 'Down conversions' of information from meaning to symbols must be adequately recovered by 'up conversions'. Commanders also require the means for checking that their sensors are working and are being used correctly.

The conclusions are intuitively sound and could be arrived at by other paths. Yet our approach has features that are unobvious: First, in modeling a command center as a Chinese Room, we had to consider how symbolic representations of battlespaces arise in practice. This led us to distinguish object-inferencing sensors from extended human sensing. Second, the idea that commanders would transact information mindlessly as John-in-the-room seemed unreasonable at first glance. But closer scrutiny suggested that it could occur (perhaps for short durations, and unintentionally), as a consequence of the 'down conversion' / 'up conversion' of information between meaning and symbols.

Our findings provide a moral basis for designing systems so that the humans work at the level of meaning. Thus we have a moral case for fostering situation awareness to levels that are qualitatively high (Endsley, 1995), and for 'story generation strategies' (Klein, 1996). Of course, this article has only established that a hazard exists to commanders (namely that their moral agency can be undermined), as a deduction from normative ethics. It is another, quite separate matter to prove that certain measures will address the hazard (repair and preserve their moral agency), in isolation or in total. As we lack a theory of treatment, we revert to highlighting the things that should be avoided, or for which treatments need to be developed.

We recommend that *ethicists should establish a mechanism that tracks the potential moral implications of choices in a system's design and intended operation*. In this article we have identified two choices that can undermine a commander's moral agency: the use of object-inferencing sensors without means for validating them, and the 'down conversion' of information from meaning to symbols. We can think of taking a Moral View of a system, tracking the moral consequences that should be addressed. The idea follows the development of Human Views, developed by ergonomists to track the implications for people as a consequence of a design (Bruseberg, 2008).

Second, we recommend that those who authorize the design of command and control systems *acknowledge that if a commander's moral agency is undermined, then the apportionings of responsibility are unclear*. There appears to be a gap in normative ethics, in that we have ways to deny moral agency, but not to affirm it. This article constructed a method for denying moral agency: we pose some necessary

condition(s), and then look for failure. Affirmation would appear to be much harder, in that we need to pose sufficient conditions. Thus the Chinese Room Argument describes a person (John) who's moral agency has been suppressed. The conditions that restore and affirm his/her moral agency are unobvious.

We emphasize that said gap in normative ethics was *not* created by this article. We have merely identified its existence. Indeed we are confronting a new notion that could be taken up in future research: that *for certain kinds of systems, it is desirable that at any time, we can articulate the agents that are morally responsible for the system's actions and the apportioning of that responsibility*. Tentatively, we might say that the system is *accountable* at all times. Said accountability might be articulated by applying the reading of Aristotle's 'compulsory' that we advanced in this article (perhaps under the analysis scheme used by (Hew, 2014)).

We might then propose some conditions under which responsibility is divisible and thus apportionable, and an algorithm for dividing up that responsibility. As part of constructing said algorithm, we could posit conditions for one agent being more responsible than another, based on how those agents interact with the system. Two features are striking: a given agent may only be able to inspect and/or affect selected sub-systems within the overall system, and the number of interventions that they could make over a given time interval may differ from another agent's. If the algorithm and its supporting conditions can be rationalized in terms of existing normative theories then well and good, otherwise we would be postulating new tenets.

Acknowledgments

The author thanks Tony Dekker, John O'Neill, Darryn Reid, Paul Whitbread and the anonymous reviewers for their feedback and suggestions. This article is UNCLASSIFIED and approved for public release. Any opinions in this document are those of the author alone, and do not necessarily represent those of the Australian Department of Defence. This article was originally published in Ethics and Information Technology. The final publication is available at Springer via https://doi.org/10.1007/s10676-016-9408-y .

Compliance with Ethical Standards

The research in this article did not involve human participants or animals.

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology, 7*(3), 149-155. doi: 10.1007/s10676-006-0004-4

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251-261. doi: 10.1080/09528130050111428

Aristotle (translated by W. D. Ross). ( ). *Nicomachean Ethics*

Bringsjord, S., Bello, P., & Ferrucci, D. (2001). Creativity, the Turing Test, and the (Better) Lovelace Test. *Minds and Machines, 11*(1), 3-27.

Bruseberg, A. (2008). Human Views for MODAF as a Bridge Between Human Factors Integration and Systems Engineering. *Journal of Cognitive Engineering and Decision Making, 2*(3), 220-248. doi: 10.1518/155534308x377090

Cole, D. (2015). The Chinese Room Argument. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition ed.).

Department of Defense (U.S.). (2011). *Joint Operations*. (Joint Publication 3-0). U.S. Department of Defense.

Department of Defense (U.S.). (2013). *Department of Defense Dictionary of Military and Associated Terms*. (Joint Publication 1-02). U.S. Department of Defense Retrieved from http://www.dtic.mil/doctrine/dod_dictionary/.

Department of Defense (U.S.). (2014). *MIL-STD-2525D Joint Military Symbology*.

Dotterway, K. A. (1988). *Systematic Analysis of Complex Dynamic Systems: The Case of the USS Vincennes.* (Masters Thesis), Naval Postgraduate School.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 32-64. doi: 10.1518/001872095779049543

Eshleman, A. (1999). Moral Responsibility. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2009 Edition ed.). Retrieved from http://plato.stanford.edu/archives/win2009/entries/moral-responsibility/.

Falcon, A. (2015). Aristotle on Causality. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Spring 2015 Edition ed.). Retrieved from http://plato.stanford.edu/archives/spr2015/entries/aristotle-causality/.

Fogarty, W. M. (1988). *Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988*. (AD-A203 577). Department of Defense.

Haksar, V. (1998). Moral agents. In E. Craig (Ed.), Routledge Encyclopedia of Philosophy (Vol. 6, pp. 499-504). London: Routledge.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D, 42*, 335-346.

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology, 16*(3), 197-206. doi: 10.1007/s10676-014-9345-6

Himma, K. (2009). Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19-29. doi: 10.1007/s10676-008-9167-5

Klein, G. (1996). The Effect of Acute Stressors on Decision Making. In J. E. Driskell & E. Salas (Eds.), *Stress and Human Performance*. Malwah, New Jersey: Lawrence Erlbaum Associates.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *Systems, Man and Cybernetics,*

Part A: Systems and Humans, IEEE Transactions on, 30(3), 286-297. doi: 10.1109/3468.844354

Polmar, N. (2001). *The Naval Institute guide to the ships and aircraft of the U.S. fleet* (Vol. 17): Naval Institute Press.

Roberts, N. C., & Dotterway, K. A. (1995). The Vincennes Incident: Another player on the stage?*. *Defense Analysis, 11*(1), 31-45. doi: 10.1080/07430179508405642

Rogers, W. C. I., Rogers, S., & Gregston, G. (1992). *Storm Center: A Personal Account of Tragedy & Terrorism*: Naval Institute Press.

Royakkers, L., & van Est, R. (2010). The cubicle warrior: the marionette of digitalized warfare. *Ethics and Information Technology, 12*(3), 289-296. doi: 10.1007/s10676-010-9240-8

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-457.

Soccio, D. J. (2009). *Archetypes of Wisdom: An Introduction to Philosophy* (7 ed.): Cengage Learning.

Williams, G. (2014). Responsibility. In J. Fieser & B. Dowden (Eds.), The Internet Encyclopedia of Philosophy. Retrieved from http://www.iep.utm.edu/.

Figure 1: Crew members monitor radar screens in the Combat Information Center aboard the *USS Vincennes*, 1 January 1988 (Photo: Tim Masterson, United States Navy)